DOI：10.13998/j.cnki.issn1002-1248.23-0251

深度学习语言模型的研究综述

王思丽[1]，张　伶[2]，杨　恒[1]，刘　巍[1]
（1. 中国科学院西北生态环境资源研究院 文献情报中心，兰州 730000；2. 新乡医学院 管理学院，新乡 453003）

摘　要：[目的 / 意义] 深度学习语言模型是当前提高机器语言智能的主要方法之一，已成为数据资源自动处理分析与知识情报智能挖掘计算不可或缺的重要技术手段，但在图情领域利用其进行技术开发和应用服务仍存在着一些困难。本研究通过系统梳理与揭示深度学习语言模型的研究进展、技术原理与应用开发方法，以期为图书馆员及同行从业者深入理解与应用深度学习语言模型提供理论依据与方法路径。[方法 / 过程] 系统地调研和梳理了深度学习语言模型的产生背景、基础性特征表示算法、代表性应用开发工具，揭示其演化发展的动态历程及技术原理，分析各算法模型与开发工具的优缺点与适用性；深入地归纳总结了深度学习语言模型应用开发面临的挑战问题，提出两种拓展其应用能力的方法策略。[结果 / 结论] 深度学习语言模型应用开发面临的重要挑战包括参数繁多，精度难调；依赖于大量准确的训练数据，变化困难；可能引发知识产权和信息安全问题等。未来可考虑从面向特定领域和特征工程两方面入手以拓展和提升其应用能力。

关键词：深度学习；语言模型；神经网络；预训练模型；词嵌入

中图分类号：G202；G250.73；TP391　　文献标识码：A　　文章编号：1002-1248（2023）08-0004-15

引用本文：王思丽, 张伶, 杨恒, 等. 深度学习语言模型的研究综述[J]. 农业图书情报学报, 2023, 35(8)：4-18.

1 引　言

深度学习是当前人工智能和机器学习领域的热点研究方向，已成为互联网数字科技行业占领行业制高点的决胜因素。对于自然语言处理、计算机视觉等的诸多任务而言，如文本分类、情感分析、机器翻译、图像 / 语音识别等，深度学习已发挥出了巨大作用且方兴未艾，未来也必将成为图情领域进行数据资源自动处理分析与知识情报智能挖掘计算不可或缺的重要技术手段。以此为认知基础，本文系统地调研和梳理了深度学习语言模型的产生背景、基础性特征表示算法、代表性应用开发工具等，揭示其演化发展的动态历程及技术原理，分析各算法模型与开发工具的优缺点与适用性，进而深入地归纳总结了深度学习语言模型应用开发面临的挑战问题，提出两种拓展其应用能力的

收稿日期：2023-04-20

基金项目：甘肃省哲学社会科学规划项目“基于大数据技术提升新闻媒体舆论监督能力研究”（2021YB158）；甘肃省自然科学基金“甘肃省医疗健康大数据资产管理模式与再利用机制研究”（23JRRA581）

作者简介：王思丽（1985- ），女，博士，副研究馆员，研究方向为知识发现与知识组织。张伶（1987- ），女，博士，讲师，研究方向为知识发现与知识组织。杨恒（1992- ），男，硕士，馆员，研究方向为自然语言处理与深度学习。刘巍（1980- ），男，硕士生导师，副研究馆员，研究方向为知识计算与知识挖掘

方法策略，以期为图书馆员及同行从业者深入理解与应用深度学习语言模型提供理论依据与方法路径。

2 深度学习语言模型的产生背景

语言和智能是人类特有的能力，如何使机器能够像人类一样进行自然语言理解和表达，进而实现一些更高层次的智能行为，如学习、思考、推理、决策等，一直以来都是人工智能的首要目标和重要挑战。在此背景下，语言模型[1]被认为是提高机器语言智能的主要方法之一，并受到学界和业界的广泛关注。在技术实现上，基于机器学习的语言模型是人类早期开发到现在仍然流行的重要方法。在机器学习方法未出现以前，不借助于人工智力，机器几乎没有任何处理未知数据问题的智能。机器学习方法通过训练大量的样本数据，根据从样本数据中学习到的知识模式，实现对未知数据问题的解答、分类与预测等。但机器学习方法已被实践证明存在很大局限性：首先，不是任何数据都能作为机器学习的样本数据，只有学习到恰当的相关的数据，机器才能预测出正确的结果，反之，则不能。但机器本身是无法判断样本数据合适与否的，也无法明确理解究竟要从样本数据中学习到什么，也就是说机器学习的输入和输出是机器自己无法控制的。其次，机器学习依赖输入的样本数据，常常需要人预先定义、从原始数据中搜集、提取、创建后提供给它。人定义、创建并提供作为机器学习输入的数据常被称之为“特征”，人从原始数据中获取、处理和生成特征的过程，又常被称为“特征工程”。实际演算时，特征即是指机器学习中的模型参数和超参数，模型参数需要从大量样本数据中学习和估计得到，而超参数需要人来设定，参数值设置的不同将会对结果预测的精确度产生很大的影响。尤其是超参数调优的过程，就是特征工程的实施过程，仍受人的主观知识和发现特征的能力所制约。此外，样本数据的分布通常具有一定的差异性和不均匀性，训练样本数据时可能需要预先做一定的假设和取舍，有的机器学习模型可能只有当未知数据符合训练时的数据分布假设时预测结果才良好，真正适用的模型需要反复训练和泛化。即便是当前训练良好的模型，在不同的应用情境下可能依旧无法做出令人满意的决策。因此，机器学习通常在相对专业、极具目标性、解空间有限的领域内能够取得巨大成功，如国际象棋、日本象棋等。目前，机器学习方法已经有大量的算法模型，不同算法模型的精确度可能大有不同，但当精确度达到一定饱和后，最终决定机器学习算法模型优良程度上限的仍是数据和特征，是人发现和抽象特征的能力。机器学习无法自主完成特征工程但又严重依赖于特征工程这一问题，使得特征工程常被认为是阻碍机器学习实现人工智能的一个瓶颈。深度学习的出现正是为了解决上述难题。如图 1 所示，深度学习与其他机器学习方法间有着明显的区别和联系。

深度学习语言模型实际上是对传统机器学习方法中神经网络算法模型的一种扩展和改进。传统神经网络算法模型常将整个多层网络整体视为一个巨大的单一的神经网络进行训练和学习，对于训练计算中出现的误差，模型只能将误差从输出层直接再反向传递回输入层，通过调整整个网络的参数来优化算法。当网络具有多层复杂结构时，每次反向传递的误差可能会逐渐缩小乃至最后消失[1]，使得顶层输入层难以获取到正确的误差反馈，也就无法对整个网络进行有效的参数调整和优化，最终造成算法的学习效果难以理想，预测的精确度变低。深度学习语言模型成功的关键是将网络深层化多层化，让每一层都参与到相应阶段的训练和学习中来，将上一层的输出数据作为下一层的输入数据，由浅入深由易入难地逐步完成学习。由于每一层都参与了学习，误差反馈可在每一层上得到及时处理，且每一层学习也可根据实际情况采用不同的学习方法。最终根据此方式进行预训练，机器将可自动由浅层的初级简单特征逐步学习到深层的高级复杂特征。目前，凭借深度学习语言模型技术机器已经能够实现对海量非结构化数据进行自动分析提炼和挖掘识别出重要特征，依靠自身能力获取恰当知识高效地完成一定的知识表示、理解、推理、解答与决策任务[2-4]。

DOI：10.13998/j.cnki.issn1002-1248.23-0251

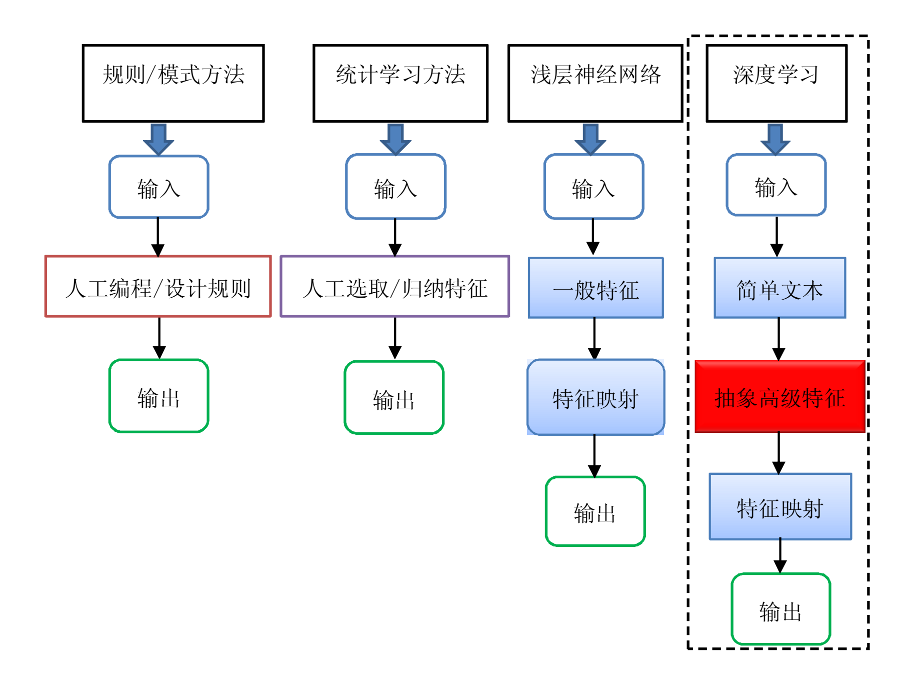


图 1　深度学习与其他机器学习方法的区别与联系

Fig.1　Differences and connections between deep learning and other machine learning methods

3　深度学习语言模型的特征表示算法

在深度学习未出现以前，学界常用的语言特征表示模型有布尔逻辑模型、向量空间模型、独热表示模型、LDA 主题模型、N-gram 统计语言模型、分布式神经网络语言模型 NNLM 等，但这些模型都存在着不同程度的局限性，特征表示和学习能力有限，准确度不高。深度学习语言模型的特征表示方式和 NNLM 类似，不同的是，它进行特征学习和表示时不仅可以考虑文本上下文的语义，而且可以使用带时间序列的数据进行训练，且对已有复杂算法进行了良好封装，简化了特征模型的构造和优化过程。深度学习在原来只有输入层和输出层的神经网络模型中增加了多个隐藏层的深度神经网络，基本流程一般包括预训练和微调两个步骤，模型的参数会在预训练阶段逐层进行学习，在微调阶段作为单个神经网络进行调优。不同深度学习语言特征表示算法的差异主要就是预训练和微调方法的不同，但也和机器学习算法模型一样分为非监督和监督学习两种。非监督学习是指训练数据是没有经过特殊处理的原始数据集，目标是尽可能地保留原始数据的特征并降低特征的维度，监督学习是指训练数据是经过人工或机器标注的数据集，目标是尽可能减小特征分类的错误率，实质都是寻求最佳的特征表达方式。下面本文着重对几种基础的具有代表性的深度学习语言模型的特征表示算法进行分析。

3.1　深度置信网络算法

深度置信网络 DBN[5]最早是在 2006 年由加拿大多伦多大学的 HINTON 等提出的，但真正作为深度学习的基础性方向开始快速发展并被广泛研究应用是在 2012 年左右。DBN 不仅可以用于特征表示和数据分类，还可以用于生成训练数据。DBN 的核心思想是构建一个观察数据（原始数据）和标签数据（标注数据）之间联合分布的概率生成模型，通过训练和调节神经元之间的权重，让整个神经网络按最大概率来生成训练数据。DBN 的预训练方法常被称为限制玻尔兹曼机（Restricted Boltzmann Machine，RBM），DBN 实质上是一个由多个 RBM 串联构成的神经网络，而每一个 RBM 实际上又是一个受限制的二值化的无向图模型：该图被限制为一个可视层和一个隐藏层，可视层的神经元主要用于接受输入与输出，隐藏层的神经元用于

提取特征，即捕捉可视层表现出来的数据相关性，不同层的神经元间可以存在链接，但同一层的神经元间不能有链接。且可视层与隐藏层的神经元具有相互独立的随机状态，取值范围一般为{0,1}，整个网络的状态将由各个神经元所对应状态的总和来确定。

单个 RBM 的训练过程，实际上是寻求一个训练样本的最大概率分布，分布的决定性因素又常取决于权重 W，因此训练 RBM 的终极目标就是寻找最佳权重 W。因而 DBN 的特征学习过程可视为一个使用逻辑回归等贪心算法逐层训练 RBM 以获取最优权重的过程。首先训练第一个 RBM（最底层 RBM），用数据向量 v 推断出隐藏层 h，获取最佳权重 W，并将 W 及隐藏层神经元的状态固定下来；然后将第一个 RBM 的隐藏层 h 作为第二个 RBM 的输入向量 v，训练出第二个 RBM 累加在第一个 RBM 上方，重复上述训练过程多次，直到获得最顶层 RBM。如果训练数据集包含带标签的数据，在最顶层 RBM 训练时，还需要将代表分类标签的神经元向量添加到输入向量 v 中共同进行训练。

3.2 卷积神经网络算法

以往的机器学习算法中，允许接收输入数据的维度基本都是一维，然而在现实应用中，数据并非总是一维，尽管可以把二维数据转化为一维输入，但会不得不丢弃很多有用信息，如时间信息，位置信息等。卷积神经网络（Convolutional Neural Networks，CNN）[6]的提出正是为了专门用来处理具有类似二维或更高维度网格结构数据的神经网络。如带时间轴的时间序列数据，可视为二维像素网格的图像数据等。在 CNN 中，输入层可以接收和处理多维数据，如经过标准化归一化的处理二维、三维或四维数组等；隐含层又包含卷积层、池化层（亚采样）、全连接层（多层感知器）等多层架构，主要通过使用卷积和池化等数学运算进行特征筛选与提取。因此 CNN 的一般训练流程为：输入 - 卷积 - 亚采样 - 卷积 - 亚采样 - 全连接 - 输出，卷积和亚采样的过程可根据实际需求重复迭代多次。

其中卷积就是指在卷积层作卷积运算，目的是为了对输入数据进行特征提取。用作卷积运算的函数常被称为卷积核，一个卷积层通常会包含多个卷积核，核中的每个元素都有一个权重值和偏差向量，类似于其他深度神经网络中的神经元。工作时，CNN 会将输入数据划分为多个区域，从每个区域中对输入数据进行特征扫描，实质是求每个卷积核内元素构成的矩阵乘积和偏差量叠加在一起的和。卷积层参数包括卷积核大小、步长和填充，三者共同决定了卷积层输出特征的大小，是 CNN 的超参数。其中卷积核大小可以指定为小于输入数据大小的任意值，卷积核越大，可提取的输入特征越复杂。此外，有的 CNN 算法在卷积层中还会使用一些激励函数以辅助提取复杂特征。在卷积层进行特征提取后，输出的特征会被进一步传递至池化层进行特征筛选和过滤，即所谓的亚采样。池化层一般包含一个预置的池化函数，其功能是将特征数据中单个点的数据替换为其相邻区域的特征数据的统计总量。池化层选取池化区域与卷积核扫描特征的方式类似，也由池化大小、步长和填充等参数控制。全接连层一般建立在 CNN 隐含层的最后面，用于向其他连接层传递信号，在全连接层中特征数据会由原来的三维结构转换为向量并通过激励函数传递至下一层。输出层一般紧邻着全连接层，使用逻辑回归函数或归一化函数输出最终的分类标签。

3.3 递归神经网络算法

一般深度学习方法的多层感知机在识别个体案例与处理一般分类任务上效果良好，但难以分析输入数据的整体逻辑序列，如具有复杂时间关联性的时间序列数据，信息内容长度多样性的结构序列数据等。递归神经网络（Recurrent Neural Network，RNN）[7]，又常被称为循环神经网络，正是为了解决带序列结构的数据问题而提出，是能够传递上下文信息的深度学习模型之一，可以处理如树、图等此类复杂的具有足够多层和节点的递归结构、拓扑结构等。RNN 可以把一个树或图结构信息编码为向量并映射到一个语义向量空间中，使得编码后语义越相似的向量距离越近。但与传统神经网络的最大不同是，RNN 的隐藏层与时间存在加权链接并构成一个循环，使得输入层与来自时间序列中上一个隐

DOI：10.13998/j.cnki.issn1002-1248.23-0251

藏层的信息将共同作用于当前隐藏层，能够在处理新输入数据的同时保存历史数据状态。如 t-1 时刻的输入数据激活的是 t-1 时刻对应的隐藏层，这些数据会被保存并在 t 时刻被传递至 t 时刻对应的隐藏层。

由于 RNN 在训练时需要考虑时间相关性及上下文信息，其训练算法也跟常规的神经网络反向传播算法 BP 不同，是一种时序反向传播算法（Back Propagation Through Time，BPTT）。在 BPTT 中，参数误差及梯度都会反向传播给时间序列的前序层，通过累加时间序列中每个元素所累积误差的权重来更新模型权重矩阵进行训练。但在实际应用中，需要将时间长度设定在有限范围内以简化训练计算的复杂度，否则会导致“梯度消失”或“梯度爆炸”。相关研究者还提出了改进版的递归神经网络语言模型 RNNLM[8,9]，可以适应更加广泛的上下文，且训练生成的单词向量可以反映单词的含义。如实现基于词向量的单词含义推理，单词向量（“king”）- 单词向量（“man”）+ 单词向量（“woman”）= 单词向量（“queen”），但其上下文长度仍受限于 N 元语法。

3.4 长短期记忆网络算法

使用常规的 RNN 训练深度学习模型常需要截短时间长度，因此难以完整的依赖时间和反映完整的上下文。虽然经研究设置 RNN 时间链接的上限可以缓解梯度爆炸问题，但梯度消失问题仍难以解决。长短期记忆网络（Long Short Term Memory Networks，LSTM）[10]正是为了解决 RNN 对时间的长期依赖问题，尤其是梯度消失问题而提出，实质是一种带时间的特殊的 RNN，主要用于处理时间序列中延迟或间隔时间相对较长的事件上下文。常规 RNN 的隐藏层只有一个状态 h，用于保存短期的信息状态，LSTM 是在原来基础上增加了一个新的单元状态 c，又被称为常量误差传送带（Constant Error Carousel，CEC），用来长期保存输入数据的值和梯度。即在 t 时刻，LSTM 的输入会有 3 个：t 时刻的输入值 x_t，t-1 时刻的输出值 h_{t-1}，t-1 时刻的单元状态 c_{t-1}。LSTM 实现的关键就是其控制单元状态 c 的方法，又被称为 LSTM 记忆模块，主要包含了 3 个被称为门的控制开关：输入门、遗忘门、输出门。

LSTM 在训练时，一般设定门的激活函数为 σ（sigmoid 函数，值域为 0 到 1），用于决定单元状态 c 中需要输出的部分，然后将 c 输入到输出的激活函数 TH（tanh 函数，值域为 -1 到 1）中，获得最终的输出值。目前 LSTM 在机器翻译、为图像生成标题、语音识别等许多应用中已获得良好精度，有关 LSTM 的变体和改进算法也已被陆续研究和提出。如 CHO 等在 2014 年提出的 GRU 算法[11]，将 LSTM 中的输入门与遗忘门改进为一个更新门来控制单元状态 c，简化了 LSTM 的计算和模型表达能力，得到了广泛认可与应用。在 GRU 被提出以后，LSTM 和 GRU 就取代了 RNN 成为常规深度学习的主流算法。

综上所述，上文这些算法可视为深度学习语言特征表示算法模型的根基。近年来，在自然语言处理领域，国内外学界又陆续提出了多种深度学习语言特征表示算法模型，比较知名的如双向长短期记忆网络 BLSTM、卷积与递归联合神经网络 CNN-RNN、双向长短期记忆与卷积联合网络 BLSTM-CNN、深度递归神经网络 DRNN、文本卷积神经网络 TextCNN[12]、快速文本分类网络 FastText[13]、文本递归神经网络 TextRNN[14]、文本递归与卷积联合神经网络 TextRCNN[15]、深度金字塔卷积神经网络 DPCNN[16]、多语言分层注意力网络 MHAN[17]、多标签注意力卷积神经网络 AttnConvnet[18]等。但万变不离其宗，这些新的算法模型仍是以经典算法模型的理念和技术为根基，经融合、扩展、改进等演化发展起来的，可视为经典算法模型的变体。这即是一种发展趋势，也是一种发展瓶颈。这意味着，当前深度学习语言模型的基础算法已经达到了一定瓶颈，亟需新的理念和技术的突破。

4 深度学习语言模型的应用开发工具

目前，业界已经提供了相对良好的工具环境支持基础性深度学习语言特征表示算法模型的应用开发，使得我们付出较小的成本代价即可快速实现一些常见深度学习语言模型的构建与训练。如谷歌大脑在 2015

年推出了开源深度学习系统框架 TensorFlow[19]，支持各类深度学习算法模型的编程实现。TensorFlow 前身是谷歌的神经网络算法库 DistBelief，实质是一个基于数据流图进行高性能数值计算的开源 API 和软件库，支持多种编程语言下的调用和开发，如 Python、C++ 等，但应用比较多的为 Python。TensorFlow 开发一般分为数据流图的定义 / 构建和图的执行 / 运算两个阶段，在第一个阶段，借助于 TensorFlow 框架的 API，能够快速构建和训练出基于复杂神经网络算法和反向传播过程的深度学习图模型；在第二个阶段，对图模型中预定义好的运算进行执行，运算中可操作的核心数据单位被称为张量（Tensor，*N* 阶矩阵），实质就是执行一个巨大的矩阵数学运算。TensorFlow 通过使用图模型将所有的可运算数据转化为图的节点，针对不同的节点可按需执行运算，可有效获取图中间某些节点的值以进行其他运算，可分配给多个 CPU 和 GPU同时执行运算等，极大地节约了系统开销和提高了执行效率，因而是目前应用最多最流行的开源深度学习框架。虽然基于 Python 的深度学习开发比较热门，但仍有很大一部分应用系统源自 Java，也急需 Java 的解决方案。Deeplearning4j（DL4J）[20]是一个由美国商业智能软件公司 Skymind 发布的专为 Java 编写的开源深度学习库，支持上述多种深度学习算法模型的 Java 实现与优化，还可与 Hadoop、Spark 集成，支持分布式运行计算等。2017 年，美国著名社交网络公司 Facebook AI 发布了基于 Python 的具有动态图模式和分布式训练性能的深度学习张量库 PyTorch[21]，由于其设计理念比较先进，一经推出就受到热烈关注继而迅速流行起来。此外，还有基于 Python 的支持自动梯度函数计算的 Theano[22]、具有高度模块化神经网络 API 的 Keras[23]、基于 C++ 的以轻量快捷著称的 Caffe、中国百度公司发布的 PaddlePaddle 平行分布式深度学习框架等。尽管相关开源框架很多，但目前受众最多的仍是基于 Python 的 TensorFlow、PyTroch 等。

但这些开源工具框架仅仅是提供了一个基础开发环境和工具平台，在使用时仍需要通过编程的方式从原始研究开始一步步自行构建相应的算法模型并进行训练和测试，导致实际应用开发难度较大。为了解决这个问题，业界又先后推出了多种深度学习语言模型的开源工具包，支持深度学习语言模型的一站式生成、预训练和微调等，使得用户不必理解算法技术原理即可实现开箱即用。下面本文着重对几种主流的具有代表性的深度学习语言模型的应用开发工具进行分析。

4.1 以 Word2Vec 为代表的词嵌入模型生成工具

Word2Vec[24]是谷歌在 2013 年发布的一个用于生成词嵌入的开源工具包，主要功能是对文本进行训练学习并转化为词嵌入模型。基于词嵌入模型，文本中的每一个词最终都会被映射到一个特定向量，词间关系的衡量转变为词向量之间距离的计算，文本的主题关系表达转变为基于词向量的 K-means 聚类等。Word2Vec 主要包含两个分类模型：CBOW 和 Skip-Gram 模型。CBOW 模型是利用目标特征词的上下文语境预测和计算出该特征词的词嵌入，训练的目标是使得在给定上下文且考虑权重的条件下获得目标词作为输出的条件概率达到最大化。CBOW 模型的实质是移除了原有 NNLM 模型中非线性的隐藏层，将所有输入的词向量都集中在同一个嵌入层，并将嵌入层与输出层直接相连接，等同于将词袋模型与一个嵌入矩阵相乘，从而得到一个连续的词嵌入。Skim-Gram 模型正好相反，是利用目标特征词预测和计算出该特征词的上下文词嵌入，训练的目标是最大限度的减少输出层上下文词嵌入预测的错误概率。Skim-Gram 模型的实质是计算输入词的词嵌入与目标词的词嵌入之间的余弦相似度，并进行归一化函数计算。同时，Word2Vec 还提供了两种学习优化算法：分层归一化算法和负采样算法。分层归一化算法的基本思想是通过构造基于哈夫曼编码的二叉树将对 *N* 个词的复杂归一化概率问题分解转化为 LogN 个词的条件概率乘积，该算法分类训练使用的负例是二叉树的其他非最优路径。负采样算法是为了解决训练样本的中心词很偏僻不适合用哈夫曼树进行遍历学习的情况，通过对样本正例进行随机负采样，建立一个正例和 *N* 个负例之间的二元逻辑回归的似然

DOI：10.13998/j.cnki.issn1002-1248.23-0251

函数进行参数求解。Word2Vec 已提供了上述训练模式的开源实现，具体应用时可通过系统提供的超参数进行不同算法模型组配调用。Word2Vec 支持多种编程语言，具有 C++、Python、Java 版本等，安装部署简单，支持在百万级以上文本数据集上进行长时高效训练，已成为目前应用最广泛最便捷的词嵌入生成工具。

随后，以 Word2Vec 为基础的扩展工具也陆续被提出。2014 年，LE 等提出了 Doc2Vec（Sentence2Vec、Paragraph2Vec）[25]，基于非监督式算法学习训练从文本中自动生成句子 / 段落 / 文档的向量模型。与 Word2Vec 类似，该模型可通过计算距离来衡量句子 / 段落 / 文档之间的相似性。同年，JEFFREY 等提出了 GloVe[26]，是一个基于全局词频统计的词嵌入生成工具。与 Word2Vec 不同的是，GloVe 模型并没有使用神经网络相关算法，而是通过对来自语料库的聚合全局词 - 词共现进行训练，构造了一个共现概率矩阵 M（矩阵中每一个元素 m_{xy} 代表单词 x 和上下文单词 y 在特定大小上下文窗口内共同出现的次数）来挖掘和表示表示词嵌入空间的线性子结构关系，然后对词嵌入和共现矩阵间的近似关系进行加权计算，构造一个损失函数实现对模型的参数求解。GloVe 与 Word2Vec 相比，能够充分利用所有语料，但计算代价和开销也比较大，因而相对应用不是特别广泛。2015 年，NIU 等提出了 Topic2Vec[27]，基于 Word2Vec，将主题结合到 NNLM 模型中，用于在与单词相同的语义空间中学习主题的分布式表示。2016 年，CHRISTOPHER 提出了 Lda2Vec[28]，将 Word2Vec 和 LDA 有机结合起来，在 Word2Vec 的 Skip-Gram 模型上进行 LDA 主题建模，使用上下文嵌入来预测上下文特征词。上下文嵌入被定义为单词嵌入和文档嵌入的总和，其中单词嵌入由 Word2Vec 生成，而文档嵌入是由文档权重向量和主题矩阵的加权组合。Lda2Vec 不仅能学习到词嵌入和上下文嵌入，还能学习到文档的主题特征表示。

4.2 以 BERT 为代表的预训练语言模型开源框架

深度学习预训练语言模型是指可以在某个数据集上训练出一个基准模型，然后只需要直接调用或微调该模型即可在其他数据集上实现各种预设功能，这一过程又常被称为“转移学习”策略。转移学习的突破是深度学习得以迅速发展的主要原因，一方面它有效解决了随着网络的不断加深和数据集的不断扩大造成的完全重新训练一个模型所需要的成本也在不断增加的问题；另一方面也非常有利于帮助那些没有时间或资源从头开始学习或构建模型的研究人员快速学习掌握相关技术。2015 年，微软研究院的 HE 等提出了深度残差网络模型[29]，率先利用残差的方式将 CNN 扩展到 100 层以上，刷新了当时最高的网络深度纪录。随后，在自然语言处理、图像识别、计算机视觉领域，采用预训练好的大型神经网络模型来提取特征以提高后续任务处理能力已成为一种常规做法。

但一般的模型通常都是基于无监督的浅层神经网络进行训练，虽然在词的等级及聚类上有着良好的特性，但却非常缺乏对连续文本的内在语义联系和上下文语言结构的良好表达能力。2017 年，谷歌首先提出了 Transformer 模型[30]，舍弃了 RNN 的循环式网络结构，采用了一种全新的注意力机制，基于固定长度的上下文来实现，可直接模拟和表达文本句子中所有单词间的关系而无需理会其各自位置。Transformer 注意力机制的基本思想是计算文本句子中的每个词与所有词的相关关系，利用相关关系来调整词的权重并获得新的词汇特征。Transformer 模型最终通过对输入文本不断进行上述的注意力机制层和一般的非线性层叠加训练来获得全局的文本语义表达。因而与常规的 RNN 和 CNN 相比，Transformer 模型在性能上要更好，训练模型所需的计算资源也更少。2018 年，美国华盛顿大学的 PETERS 等提出了 ELMo 模型[31]，是一种将向量和嵌入结合起来表示单词的新方法，主要基于双层双向 RNN 或 LSTM 进行计算生成词嵌入。与以往的词嵌入生成工具的最大不同是，可以考虑词嵌入的完整输入语句，将表征作为特征传递给下游任务，使得相同单词在不同的上下文语境中具有不同的词嵌入。紧接着，大型深度学习网站 Fastai 和人工智能企业 DeepMind 联合推出了 ULMFiT 模型[32]。该模型可对预训练

语言模型进行微调并针对每一层设置不同的学习率，将其在维基百科的长期依赖建模数据集之一的 WikiText-103 上进行训练，从而得到新的数据集并且不会忘记之前学习过的内容特性。使用 ULMFiT 模型，只需要较少的数据集就能产生比一般文本分类模型更好的效果。随后，美国 OpenAI 公司发布了 GPT-2 模型[33]，该模型实质是一个具有 12 层 Transformer 结构的 Transformer 模型，其预训练语言模型是基于百度 15亿词汇文本和 800 万 Web 数据集进行训练的一个单向语言模型。

2018 年 11 月，谷歌发布了重量级开源框架 BERT 模型[34]，沿袭了 GPT 模型的基本架构，采用 Transformer 编码器作为主体结构，使用纯文本语料进行训练。BERT 使用的训练数据是涵盖了约 33 亿词汇的开源语料库 BooksCropus 及英文维基百科数据，BERT 模型标准版约有 1 亿参数量（与 GPT 模型大致相当）。但与 GPT 模型或其他只考虑词的单侧上下文的模型不同，BERT 可以同时考虑词的两侧，并进行多任务学习，是首个无监督的支持双向深度预训练的双向语言模型。此外，BERT 还具有许多其他创新特性，如可以采用遮蔽词（MaskLM）方式来标记训练，可以进行句子级别的连续性预测任务等。BERT 模型一经发布即获得最高热度关注，已在多个 NLP 任务上取得惊人效果。2019 年初，谷歌又发布了 Transformer-XL 模型[35]，作为 Transformer 模型的改进版，可以帮助机器理解超出固定长度限制的上下文，极大的提高了模型的灵活性和推理速度，已在多个语言建模基准数据集上都取得了新的进展。2019 年 6 月，谷歌大脑和美国卡耐基梅隆大学联合推出了 XLNet 模型[36]，借鉴了 Transformer-XL 模型中当前最先进的自回归理念，其一，利用自回归方法解决了 BERT 模型中存在的局限性问题，如 BERT 忽略了被遮蔽位置之间的依赖关系，模型存在预训练 - 微调差异等；其二，通过最大化模型中因子分解顺序所有排列可能的期望值实现了双向上下文信息的学习；此外，还将 Transformer-XL 的分段重复机制和相对编码方案集成到了预训练过程中，改进了文本处理性能。XLNet 实质是一种基于广义置换语言建模目标的无监督表示学习方法，是一种泛化的自回归预训练模型，目前测试表明，其在长文本语言表示任务上性能显著且优于 BERT，如自动问答、情感分析、自然语言推理、文本分类等。由于 BERT 和 XLNet 显示的良好效应，国内外相关机构团队通过改进和扩充其预训练任务、语料和时间等，先后生成了一批覆盖更多领域场景数据和任务的中文 BERT 模型，如清华大学推出的百度百科 BERT[37]，哈尔滨工业大学发布的 BERT-wwm[38]，美国 Facebook AI 和华盛顿大学联合发布的 RoBERTa-zh-Large 等[39]。中国科大讯飞也陆续开源了多个面向通用领域的文本识别、语义理解的中文预训练语言模型。

虽然上述预训练语言模型都已开源，但由于比较分散，目前也有一些开源自然语言处理库将多种主流预训练模型整合起来供按需调用。如德国 Zalando Research 公司发布的 Flair[40]，已将 GloVe、BERT 等多个模型集成起来供调用，并推出了命名实体识别、文本分类、训练定制模型等 NLP 服务。美国斯坦福大学开发的 StanfordNLP[41]，支持超过 53 种语言，基于 PyTorch 构建并打包了多个预训练语言模型，也包括命名实体识别、实体关系抽取、依存句法分析等。中国百度公司也推出了开箱即用、可灵活定制的 PaddleNLP 工具集，覆盖了自然语言理解与生成的多模态应用场景，提供信息抽取、文本分类、情感分析、语义检索、知识问答等多项任务的快速实践支持。

4.3 以 GPT 为代表的大规模语言模型应用程序

ChatGPT 是美国 OpenAI 公司在 2022 年 11 月推出的一款智能聊天机器人程序[42]。与以往功能简单、机械生硬的普通聊天程序或客服助手不同，它不仅能够和人类进行基本的聊天对话，而且能够深入理解和主动学习人类的语言观念、情感思维、意识形态和意图动机等，基于聊天的上下文内容信息以及针对人类提出的各种问题和提示，和人类进行连贯的互动交流和真正的协作创新，进而完成一些高难度的场景任务，如智能问答、考试答题、撰写文案、编写代码、创作

DOI：10.13998/j.cnki.issn1002-1248.23-0251

论文、翻译文本、分析数据、以文生图等。因而，ChatGPT 一经推出即受到热烈追捧，目前已成为世界上用户增长最快的应用程序，连比尔盖茨都称赞 ChatGPT 出现的意义不亚于计算机和互联网的诞生。ChatGPT 实质上是一个人工智能和深度学习技术驱动的自然语言处理程序和大规模通用语言模型，通过创建多层次的深度神经网络和可预测可扩展的深度学习栈，并嵌入了人类反馈强化学习（RLHF）和监督微调机制[43,44]，使得模型能够灵敏感知和准确理解不同语言风格和语境模式的微妙差异，然后依据应用场景进行重新组合、概率排序和模仿推导等，从而生成更具有真实性和创造性的结果。迄今为止，ChatGPT 已经经历了 GPT-1 至 GPT-4 多个版本的演化。其中，GPT-4[45]于 2023 年 3 月发布，提升了对多模态功能的支持，包括对文字、语音、图像、视频的输入和输出处理、理解优化和加强等。实验表明，ChatGPT 尤其是 GPT-4 在各种专业测试和学术基准上的表现已与人类旗鼓相当。

其实，在 ChatGPT 出现以前，国内外已经有相关机构企业发布过大规模语言模型或聊天机器人，如谷歌的 LaMDA，Meta 的 OPT-MIL、BlenderBot，Hugging Face 的 Bloom，DeepMind 的 Sparrow 等，但都反响平平。如今，以 ChatGPT 为引领，国内外多个互联网机构开始竞相投入生成式大规模语言模型及相关产品的深度研发和布局。如微软已将 ChatGPT 嵌入 Office 办公套件和 Bing 搜索引擎，谷歌发布了基于 LaMDA 模型的对话机器人 Bard，Meta 开源了 LLaMA，百度公开了“文心一言”ERNIE Bot，阿里巴巴推出了 M6-OFA，腾讯推出了“混元”系列大模型，京东推出了 ChatJD，华为联合鹏程实验室发布了“鹏程·盘古”大模型，复旦大学发布了 MOSS，IDEA 研究院发布了“封神榜”大模型等。此外，360、浪潮、快手、有道等企业也陆续宣布正在推进相关同源技术和大规模语言模型的专项研究。可见 GPT 及类似技术不仅是一种先进的强智能的生产工具，能够为各行业提供数据、算力、模型等基础人工智能服务能力，而且可能会带来一场全新的划时代的生产力革命，有望成为下一代信息产业基础设施并重构和形成新的应用生态，有效辅助各行业、各产业链获得更高的全生命周期质量、效益和核心竞争力。但遗憾的是，GPT 目前在中国并不开源，而国内现有的类 GPT 工具模型与其能力还相差较大，且更多的也是不开源，亦或者仍处于试用或保密研发阶段。因而 GPT 目前仍难以嵌入图情机构知识管理与服务系统及提供和形成便捷化的开发应用支持和普及性的智能化服务。

5 深度学习语言模型的应用开发挑战及策略

5.1 应用开发面临的挑战

深度学习作为一种跨越式的创新，一直以来国际上对其研究都十分活跃，已发布了大量的开源算法模型和工具框架供应用开发。尽管不同工具框架的功能特性和编程语言大有不同，但也具有一些通用性，如基本都已封装了一些主流深度学习语言模型并提供了调用接口，使得可以在不必深入了解算法原理的情况下也能快速构建多种深度学习模型，且模型的训练可通过定义一个深度学习层结构来实现，用户可以专注于进行参数设置与调优，而无需关心算法的具体实现。再如多数工具框架都可在不同操作系统及 CPU、GPU 或 TPU 上便捷切换，用户可以专注文本特征分析挖掘与特征工程的实现，而不需要考虑太多硬件环境。尽管已经见识到了以 GPT 为代表的生成式深度学习语言模型在数据资源自动发现获取与知识情报智能挖掘分析方面的超强能力，但在图情领域仍存在着较多观望等待和坐享其成心理，理论与实践能力不匹配，技术开发和应用服务落地困难等问题。究其原因，深度学习语言模型的开发应用仍面临着一些重要挑战。

（1）深度学习语言模型的参数繁多，精度难调。与传统机器学习与浅层神经网络算法相比，深度学习算法模型中存在大量的参数和超参数，如上下文窗口的大小，隐藏层的层数，每个隐藏层神经元的数量等。训练与测试中，也需要一些特殊参数的配置，如训练样本的大小，投影学习矩阵的大小，卷积核的大小，

学习的速率，Dropout 的比率，优化算法的选择等。以 ChatGPT 为例，GPT-1 约有 1.17 亿参数，GPT-2 约有 15 亿参数，GPT-3 约有 1 750 亿参数，GPT-4 已达到约 100 万亿参数[45,46]，与人类大脑神经元的数量相当。尽管有开源模型工具作基础，但定义和构造一个深度神经网络结构的过程，还是一个对大量参数进行不断调配组合的过程，只有良好稳定合理的参数调整才能确保深度学习的有效性与提高深度学习的精度。这也意味着参数具体该如何调整如何取值，单凭经验或借鉴前人研究并不可靠，需要进行更多额外的实验才能获得。而图情领域从业者以情报研究或知识服务为主，往往很少能够熟练掌握深度学习语言模型构建与调参训练所依赖的技术开发工具与生产环境。

（2）深度学习语言模型依赖于大量准确的训练数据，变化困难。深度神经网络的结构深而复杂，对于简单的样本和问题，深度学习难以进行训练和分类预测，只有大量足够准确的训练数据，才能实现对相关权重的充分优化。并且随着数据量的增加，训练时间和成本也会不断增加。类似 ChatGPT 的大语言模型，训练或优化一次需要长达数月时间（GPT-4 为 6 个月），训练成本在 200 万美元至 1 200 万美元之间，集成用于搜索计算服务的代价更高（如果将 ChatGPT 部署到谷歌搜索引擎，粗略计算需要 51 万余台 A100 HGX 服务器和 410 万余个 A100 GPU 支持，总成本将超过 1 000 亿美元[46]），因而进行广泛的特定调整完全是不现实和不可行的。ChatGPT 的做法是建立了可预测可扩展的深度学习栈，对基础设施进行扩展，使其尽可能地在多规模场景下都具有可预测行为。此外，由于深度神经网络算法常包含了大量随机操作和 Dropout 操作[47]，再加上不同计算机的计算精度不同且有限，使得权重优化与计算的值可能会随着实现方法的不同而出现波动，因而从实验中得到的准确率很可能依赖于所使用的开源模型工具库的实现，使用不同的库进行同一种深度神经网络算法的训练，可能得到的结果也会有一定差异。目前在图情相关业务领域，经典机器学习算法仍占主导地位，如支持向量机 SVM、朴素贝叶斯 NB、条件随机场 CRF、逻辑回归 LR、随机森林 RF 等，已得到广泛认可与大量验证，且能够快速响应和实现更新等。虽然深度学习在学界和业界已经炙手可热，但在实际生产实践与服务应用中还需综合考虑应用成本、模型更新优化效率等问题，除非有足够的资本设施及保障策略支持，否则在深度学习还没有成为普惠性的人工智能服务前，其短时间内仍难以也不可能完全取代实施代价较小的机器学习方法。

（3）深度学习语言模型可能引发知识产权和信息安全等问题。随着深度学习语言模型的流行和发展，对其可能引起的知识产权、信息安全、隐私伦理和环境污染问题的关注及研究也越来越多。如模型学习能力依赖于对海量文本语料的挖掘和训练，可能对他人作品成果进行复制使用以及创作风格进行借鉴模仿从而引发新型版权侵权风险；模型生成的高度逼真的合成性内容以及高度敏感的隐私性信息（如医疗健康数据、财务状况数据、身份信息数据等）可能被用于冒充或欺骗他人从而引发隐私侵犯、电信诈骗等违法犯罪行为；模型响应可能存在政治 / 性别 / 种族偏见或歧视、违背血缘关系和伦理常识等误导性问题；模型训练与优化消耗的巨大算力可能引起碳排放问题等。相关机构团队正在研发检测和缓解这些问题的方法，如使用更加多样化的训练数据，加强透明度、问责制、审查制和知情权，推动健全人工智能应用相关政策法规等。OpenAI 于 2023 年 1 月底推出了 AI 生成内容鉴别工具[48]，旨在识别 ChatGPT 生成的文本内容，但目前仍存在较大局限性，准确率有待提高。中国相关机构也于 2023 年 3 月份联合推出了首个 AI 生成内容检测工具 AIGC-X[49]，旨在对人工智能技术生成的虚假信息、抄袭内容、垃圾邮件等进行检测，目前对中文文本的检测效果表现良好，但也存在反改写监控能力差等问题。近期权威期刊《自然》中的一篇论文也指出，ChatGPT 用于科学界必须首要遵循人类审查原则[50]。这无疑对图情及相关知识服务机构也提出了新的需求和挑战，在目前欠缺审查机制和好用工具的情况下，如何对人工智能和深度学习语言模型生成内容进行循证溯源和质量审查及控制必须引起重视。

DOI：10.13998/j.cnki.issn1002-1248.23-0251

5.2 应用能力拓展方法

综上研究，与经常需要持续更新的模型不同，深度学习语言模型可能更适合于问题比较复杂，构造模型的数据集很大且不会总要求变化的应用场景，这样一旦使用大规模数据集训练好模型便具有通用性和稳定性，能够在一定范围内长期使用。鉴于此，本研究从便于图书馆员及同行从业者利用深度学习语言模型进行知识管理决策与情报挖掘分析相关技术开发与服务应用角度出发，尝试提出了两种拓展深度学习语言模型应用能力的方法策略。

（1）面向特定领域的拓展方法。有的学科领域天生适合利用深度学习语言模型来解决问题，且领域专业数据资源丰富，可着重考虑以这些学科领域为核心，将多种前沿先进技术和领域专业数据资源强强联合起来，进行跨学科领域交叉融合研究与应用。如在面向生物医学领域进行知识服务时，可充分结合深度学习语言模型及知识图谱技术构建医学知识图谱和智能问答决策系统，对电子诊疗记录、临床试验数据、个人健康数据、医学影像等多模态医疗数据进行深度挖掘分析，以发现新的诊疗方案、预测新的疾病及评估可能出现的新的医疗风险等。总之，本方法是以特定领域的需求出发实现扩展，其综合策略如下：①领域数据的收集与准备。收集并准备与目标领域相关的大量数据，以便让模型学习到该领域的专业知识和术语。在数据准备的过程中，需要注意数据的质量和数据的平衡性，避免数据偏差或过拟合的情况。②模型架构的选择。针对不同领域，选择不同的深度学习语言模型架构，比如 Word2Vec、BERT、GPT 等。同时，也可以通过添加特定领域的知识和任务来改进模型的表现，比如 Fine-tuning、Transfer Learning 等方法。③领域专家的参与。领域专家可以为深度学习语言模型提供专业的领域知识，指导数据收集和准备，并对模型的结果进行验证和调整。通过与领域专家的密切合作，可以确保模型的应用效果更加准确和实用。④针对具体任务的优化。结合具体任务需求进行深度学习语言模型的调整和优化。比如，在文本分类任务中，可以使用分类器对模型进行进一步训练和优化，以提高分类准确率和效率。综上所述，面向特定领域拓展深度学习语言模型的应用能力需要全面考虑领域数据的收集与准备、模型架构的选择、领域专家的参与以及针对具体任务的优化等方面，以确保模型的数据来源更加可靠和安全，应用效果更加准确和实用。

（2）面向特征工程的拓展方法。特征工程是深度学习语言模型提高预测精度的决定性因素。若对原始数据（输入）和要预测的数据（输出）没有任何限制，也很难用深度学习语言模型达成某种目的。对自然语言处理来说，一般使用稀疏向量结合 N 元语法作为特征来表示单词，若不使用含数字的单词向量作为特征显然在当前技术条件下是不合理的，也无法进一步执行训练计算。最重要的是，输入特征也应当在一定时间内是固定的或有限变化的，输出的分类模式也应是有限的，深度学习无法使用连续变化的数据去预测出可能是连续变化的结果，但可以对输入和输出任务进行分解，将深度学习模型应用于其前序任务，或将模型进行适当剪裁以适应实际特征，甚至可通过调整输入和限制并细化分类结果。同时，还应掌握不同深度学习语言模型所适用的特征工程，高质量的、可伸缩性强、可解释性强的特征工程能够显著提升模型预测性能，简化模型复杂度，降低模型维护成本等。尽管，深度学习语言模型的训练一般不需要手动进行特征工程，因为它们通常可以通过对大量数据进行端到端学习，自动地学习到语言的各种特征。然而，对于一些特定的任务，仍然需要手动提取特征，以帮助深度学习语言模型更好地进行学习和应用。总之，本方法是指从通过设计输入数据或调整输入的值来适配深度学习语言模型；或通过限制输出或调整预测问题来提高模型的预测分类性能，其综合策略如下：①选择适当的特征。根据具体任务的需求选择适当的特征。比如，在舆情监测分析任务中，可以通过提取词袋模型中的特征，比如单词、词组、情感词等，来进行特征提取。②特征预处理。对于一些文本特征，如单词、字符等，可能需要进行预处理，以提高模型的性能。具体如可以通过对单词进行词干提取、词形还原等操作，以减

少单词形态的变化对模型带来的干扰。③特征选择。在一些情况下，特征过多可能会导致模型的过拟合问题。因此，需要使用特征选择技术来筛选出对目标变量影响最大的特征。比如，可以使用阈值过滤、正则化等方法对特征进行选择。④特征降维。一些高维特征可能会导致模型的运算复杂度增加，因此需要使用特征降维技术来减少特征的数量，同时保留重要的信息。比如，可以使用主成分分析 PCA、核函数等方法对高维特征进行降维处理。综上所述，面向特征工程拓展深度学习语言模型应用能力的方法策略包括选择适当的特征、特征预处理、特征选择以及特征降维等。这些方法策略可以帮助提高深度学习语言模型的性能和效率，使其更适合应用于特定的任务需求。

目前，随着越来越多的机构关注和采用类 ChatGPT 的生成式语言模型技术，已经引发了社会和市场对提示工程师（Prompt Engineer）、AI 训练师的新型职业需求，但国际上相关人才供给仍处于较为匮乏阶段。成立于 2021 年的 AI 初创公司 Anthropic 为提示工程师和图书馆员职位招聘提供了高达 17.5~33.5 万美元的年薪[51]。这也为图书馆员的未来发展提供了良好机遇与可能方向。未来，深度学习语言模型等生成式人工智能在专业 / 垂直领域应用首当其冲的问题可能是缺乏高质量的合规的标注与训练语料，而图情机构还可以作为海量高标准领域专业数据训练语料库的开发者与提供者，以保持优势地位。总的来看，人工智能和深度学习已经极大地改变了科学研究和生产实践的范式，颠覆了传统的文献情报知识发现获取、分析挖掘、组织集成与应用服务方式[52-54]，使其更加便捷化、高效化、智能化，实现了跨越式的进步。但同时也带来了新的安全风险和挑战，未来可能还会受到更多更大的冲击。

我们应该正确面对新机遇与新挑战，紧紧把握住类 GPT 等人工智能和深度学习技术带来的良好机遇和巨大红利，以深度学习语言模型相关技术为驱动力，以科研工作和社会发展需求为导向，基于图情机构已有文献数据资源和知识服务优势，开展具有自主知识产权的创新型的专业 / 垂直领域智能知识管理决策与应用服务技术及系统研发。不断加强图情领域文献数据资源与智能技术集成化研发应用能力建设，积极探索应对知识情报内容安全隐患问题的新方法和新策略，共同加快推进图情机构转型升级与创新发展，才是长远生存之道。

参考文献：

[1] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[J]. arXiv Preprint, arXiv: 2303.18223, 2023.

[2] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: A survey[J]. Science China technological sciences, 2020, 63(10): 1872-1897.

[3] 毛进, 陈子洋. 基于深度学习的科技文献摘要结构功能识别研究[J]. 农业图书情报学报, 2022, 34(3): 15-27.
MAO J, CHEN Z Y. A Deep learning based approach to structural function recognition of scientific literature abstracts[J]. Journal of library and information science in agriculture, 2022, 34(3): 15-27.

[4] 康明. 深度学习预训练语言模型-案例篇: 中文金融文本情绪分类研究[M]. 北京: 清华大学出版社, 2022.
KANG M. Deep learning pre-training language model-case: A study on emotion classification of Chinese financial texts[M]. Beijing: Tsinghua University Press, 2022.

[5] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[6] YUSUKE S. Java deep learning essentials[M]. Beijing: China Machine Press, 2017: 97-113.

[7] IENCO D, GAETANO R, INTERDONATO R, et al. Combining sentinel-1 and sentinel-2 time series via RNN for object-based land cover classification[C]// IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. Piscataway, New Jersey: IEEE, 2019: 4881-4884.

[8] JI S H, VISHWANATHAN S V N, SATISH N, et al. BlackOut: Speeding up recurrent neural network language models with very large vocabularies[J]. arXiv Preprint, arXiv: 1511.06909, 2015.

[9] RNNLM Toolkit[EB/OL]. [2023-02-20].https://github.com/IntelLabs/rnnlm.

[10] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. arXiv Preprint, arXiv: 1409.3215,

DOI：10.13998/j.cnki.issn1002-1248.23-0251

2014.

[11] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv Preprint, arXiv: 1406.1078, 2014.

[12] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv Preprint, arXiv: 1408.5882, 2014.

[13] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[J]. arXiv Preprint, arXiv: 1607.01759, 2016.

[14] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning[J]. arXiv Preprint, arXiv: 1605.05101, 2016.

[15] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for text classification [C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. New York: ACM, 2015: 2267-2273.

[16] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 562-570.

[17] PAPPAS N, POPESCU-BELIS A. Multilingual hierarchical attention networks for document classification[J]. arXiv Preprint, arXiv: 1707.00896, 2017.

[18] KIM Y, LEE H, JUNG K. Attention-based convolutional neural networks for multi-label emotion classification[EB/OL]. [2018-01-01]. http://sciencewise.info/articles/1804.00831/.

[19] TensorFlow[EB/OL]. [2023-02-25]. https://tensorflow.google.cn/.

[20] Deeplearning4j [EB/OL]. [2023-02-25]. https://github.com/deeplearning4j.

[21] PyTorch[EB/OL]. [2023-02-25]. https://pytorch.org/.

[22] Theano[EB/OL]. [2023-02-25]. https://pypi.org/project/Theano/.

[23] Keras[EB/OL]. [2023-02-25]. https://keras.io/.

[24] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint, arXiv: 1301.3781, 2014.

[25] LE Q V, MIKOLOV T. Distributed representations of sentences and documents[C]//ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing, China: ICML, 2014(32): 1188-1196.

[26] JEFFREY P, RICHARD S, CHRISTOPHER D M. GloVe: Global vectors for word representation[EB/OL]. [2018-12-29]. https://nlp.stanford.edu/projects/glove/.

[27] NIU L Q, DAI X Y, ZHANG J B, et al. Topic2Vec: Learning distributed representations of topics[C]// 2015 International Conference on Asian Language Processing(IALP). Piscataway, New Jersey: IEEE, 2016: 193-196.

[28] MOODY C E. Mixing dirichlet topic models and word embeddings to make lda2vec[J]. arXiv Preprint, arXiv: 1605.02019, 2016.

[29] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey: IEEE, 2016: 770-778.

[30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[J]. arXiv Preprint, arXiv: 1706.03762, 2017.

[31] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv Preprint, arXiv: 1802.05365, 2018.

[32] REDDY R. Universal language model fine-tuning for text classification[J]. arXiv Preprint, arXiv: 1801.06146, 2018.

[33] GPT-2[EB/OL]. [2023-02-28]. https://github.com/openai/gpt-2.

[34] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint, arXiv: 1810.04805, 2019.

[35] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: Attentive language models beyond a fixed-length context[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019.

[36] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: Generalized autoregressive pretraining for language understanding[J]. arXiv Preprint, arXiv: 1906.08237, 2019.

[37] ZHONG H X, ZHANG Z Y, LIU Z Y, et al. Open Chinese language pre-trained model zoo [EB/OL]. [2020-03-18]. https://github.com/thunlp/OpenCLaP.

[38] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word

masking for Chinese BERT [EB/OL].[2023-03-09]. https://github.com/ymcui/Chinese-BERT-wwm.
[39] XU L. RoBERTa for Chinese[EB/OL]. [2022-06-15]. https://github.com/brightmart/roberta_zh.
[40] ALAN A, DUNCAN B, ROLAND V. Contextual string embeddings for sequence labeling[EB/OL]. [2023-03-10]. https://github.com/zalandoresearch/flair.
[41] Stanford NLP[EB/OL]. [2023-03-10]. https://github.com/stanfordnlp.
[42] ChatGPT: Optimizing language models for dialogue[EB/OL]. [2023-03-16]. https://openai.com/blog/chatgpt.
[43] NISAN S, LONG O, JEFFREY W, et al. Learning to summarize with human feedback[C]//Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020: 3008-3021.
[44] LEO G, JOHN S, JACOB H. Scaling laws for reward model overoptimization[J]. arXiv Preprint, arXiv: 2210.10760, 2022.
[45] GPT-4[EB/OL]. [2023-03-16]. https://openai.com/product/gpt-4.
[46] 刘高畅, 杨然. ChatGPT 需要多少算力[R/OL]. 北京: 国盛证券, 2023.
LIU G C, YANG R. How much computing power does ChatGPT require[R/OL]. Beijing: Guosen Securities, 2023.
[47] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of machine learning research, 2014, 15: 1929-1958.
[48] AI text classifier[EB/OL].[2023-03-16]. https://platform.openai.com/ai-text-classifier.
[49] AIGC-X[EB/OL]. [2023-03-16]. http://ai.sklccc.com.
[50] VAN DIS E A M, BOLLEN J, ZUIDEMA W, et al. ChatGPT: Five priorities for research[J]. Nature, 2023, 614(7947): 224-226.
[51] Prompt engineer and librarian[EB/OL]. [2023-03-31]. https://jobs.lever.co/Anthropic/e3cde481-d446-460f-b576-93cab67bd1ed.
[52] 张智雄, 钱力, 谢靖, 等. ChatGPT 对科学研究和文献情报工作的影响[R/OL]. 北京: 国家科技图书文献中心 & 中国科学院文献情报中心, 2023.
ZHANG Z X, QIAN L, XIE J, et al. The Impact of ChatGPT on scientific research and documentation and information work [R/OL]. Beijing: National Science and Technology Library & National Science Library of Chinese Academy of Sciences, 2023.
[53] 张晓林. 从猿到人: 探索知识服务的凤凰涅槃之路[J]. 数据分析与知识发现, 2023, 7(3): 1-4.
ZHANG X L. From ape to man: Exploring the phoenix nirvana road of knowledge service [J]. Data analysis and knowledge discovery, 2023, 7(3): 1-4.
[54] 曹树金, 曹茹烨. 从 ChatGPT 看生成式 AI 对情报学研究与实践的影响[J]. 现代情报, 2023, 43(4): 3-10.
CAO S J, CAO R Y. Influence of generative AI on the research and practice of information science from the perspective of ChatGPT[J]. Journal of modern information, 2023, 43(4): 3-10.

DOI：10.13998/j.cnki.issn1002-1248.23-0251

Review of Deep Learning for Language Modeling

WANG Sili[1], ZHANG Ling[2], YANG Heng[1], LIU Wei[1]

(1. Literature and Information Center of Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000; 2. School of Management, Xinxiang Medical University, Xinxiang 453003)

Abstract: [Purpose/Significance] Deep learning for language modeling is one of the major methods and advanced technologies to enhance language intelligence of machines at present, which has become an indispensable important technical means for automatic processing and analysis of data resources, and intelligent mining of information and knowledge. However, there are still some difficulties in using deep learning for language modeling for technology development and application service in the library and information science (LIS) field. Therefore, this study systematically reviews and reveals the research progress, technical principles, and development methods of deep learning for language modeling, with the aim at providing reliable theoretical basis and feasible methodological paths for the deep understanding and application of deep learning for language modeling for librarians and fellow practitioners. [Method/Process] The data used in this study were collected from the WOS core database, CNKI literature database, arXiv preprint repository, GitHub open-source software hosting platform and the open resources on the Internet. Based on these data, this paper first systematically investigates the background, basic feature representation algorithms, and representative application development tools of deep learning for language modeling, reveals their dynamic evolution and technical principles, and analyzes the advantages and disadvantages and applicability of each algorithm model and development tool. Second, an in-depth analysis of the possible challenging problems faced by the development and application of deep learning for language modeling was performed, and two strategic approaches to expand their application capabilities were put forward. [Results/Conclusions] The important challenges faced by the application and development of deep learning for language modeling include numerous parameters and difficulties to adjust accuracy, relying on a large amount of accurate training data, difficulties in making changes, and the intellectual property and information security issues. In the future, we will start from two aspects of specific domains and feature engineering to expand and improve the application capabilities of deep learning for language modeling. Specifically, we focus on consideration of the collection and preparation of domain data, selection of model architecture, participation of domain experts, and optimization for specific tasks, in order to ensure that the data source of the model is more reliable and secure, and the application effect is more accurate and practical. Moreover, the strategic methods for feature engineering to expand the application capabilities of deep learning for language modeling include selecting appropriate features, feature pre-processing, feature selection, and feature dimensionality reduction. These strategies can help improve the performance and efficiency of deep learning for language models, making them more suitable for specific tasks or domains. To sum up, LIS institutions should leverage the deep learning for language modeling related technologies, guided by the needs of scientific research and social development, and based on advantages of existing literature data resources and knowledge services; they should carry out innovative professional or vertical domain intelligent knowledge management and application service, and develop technology and systems with independent intellectual property rights, which is their long-term sustainable development path.

Keywords: deep learning; language model; neural network; pre-trained model; word embedding